INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Calculability of the Semantics of English Nominal Compounds: Combining General Linguistic Rules and Corpus-based Semantic Information

Cécile Fabre, Pascale Sébillot

N° 2742

1995

PROGRAMME 3

ISSN 0249-6399

Calculability of the Semantics of English Nominal Compounds: Combining General Linguistic Rules and Corpus-based Semantic Information

Cécile Fabre, Pascale Sébillot

Programme 3 — Intelligence artificielle, systèmes cognitifs
et interaction homme-machine
Projet Repco

Rapport de recherche n°2742 — 1995 — 39 pages

Abstract: Our project focuses on the calculability of the semantics of nominal compounds. Our goal is to design a general model, based on domain-free lexical information, in order to exhibit and implement the principles of nominal compound interpretation. This model is based upon a precise semantic characterization of nominal constituents, which relates nouns to the predicative information that must be identified to retrieve the underlying relation of the compound. The predicate is deduced from the morpho-syntactic and semantic features of the nouns and its argument structure is used to characterize the roles of each constituent. We describe our model of interpretation of English compounds and evaluate it from the results of a program that implements this general framework. We suggest solutions to enrich this model and to adapt it to the characteristics of the compounds of a specialized corpus through the extraction of specific semantic information.

Key-words: Natural Language Processing, Nominal Compounds, Automatic Interpretation, Corpus-Based Information Retrieval

(Résumé : tsvp)

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 99 84 71 00 – Télécopie : (33) 99 84 71 71

Calculabilité de la sémantique des composés nominaux en anglais : Associer des règles linguistiques générales et des informations sémantiques acquises sur corpus

Résumé : Notre projet consiste à évaluer la calculabilité de l'interprétation des composés nominaux. Nous élaborons un modèle général basé sur des informations lexicales hors domaine afin de formuler les principes d'interprétation des composés nominaux. Ce modèle consiste à déterminer le type de relation prédicative qui est potentiellement attachée à un nom, et qui se trouve actualisée dans le cadre d'un composé. Si le nom est un déverbal, l'information prédicative est déterminée par ses traits morpho-syntaxiques. S'il s'agit d'un nom simple, nous utilisons des critères sémantiques. Dans les deux cas, la structure argumentale du prédicat sert à identifier les rôles respectifs des deux constituants. Nous décrivons notre modèle pour l'interprétation des composés anglais et nous l'évaluons à partir des résultats des tests que nous avons menés. Nous suggérons la façon dont ce modèle peut être enrichi à l'aide d'informations sémantiques extraites des textes afin d'adapter l'interprétation aux spécificités d'un corpus donné.

Mots-clé : traitement automatique du langage naturel, composés nominaux, interprétation automatique, acquisition d'informations sur corpus

Introduction

Our project[1] focuses on the calculability of the semantics of nominal compounds. These sequences, such as *system interface* in English or *réseau de distribution* in French, are very frequent in technical texts. Compounding is a productive linguistic process, which provides a way to build new denominations and to enrich the terminology of a domain by the adjunction of modifiers[2] to existing nouns or nominal sequences. For example, in the domain of computer science, one can build from the noun *generator* the terms *interface generator* or *program generator*[3]: compound nominals are used to express the sort-of relation which organizes the terminology of a domain.

Our purpose in this paper is to calculate the semantics of noun + noun (NN) sequences in English, that is to characterize and formulate the features that are needed to infer the relation between the constituents. We implement a model that predicts the semantic relation between the constituents of a compound, given their lexical description and independently of any domain knowledge. We want to avoid the problems raised by two types of interpretation systems:

- domain-dependent systems (e.g. [Mar84], [SV94]). The efficiency of such systems is limited to the domain they are built for, since they are based on very detailed lexical information written to assist the interpretation task.

[1]This project is supported by the CNET (contract CNET-INRIA n°951B030).

[2]In a noun + noun sequence, we distinguish between the head of the sequence - generally the rightmost constituent in English - and the modifier. The modifier characterizes the head in some way.

[3]The compounding process being recursive, one can make a further distinction between *parsing program generator* and *lexical analysis program generator*, and so on. In this work, we only focus on non-recursive terms, that is noun + noun sequences, because compounds with three constituents or more raise furthermore the problem of ambiguous bracketing, e.g. ((parsing program) generator) or (parsing (program generator)). Nevertheless, once the bracketing of the constituents is performed, the interpretation task may be extended to these terms. See [Don82] and [LD95].

- domain-independent systems (e.g. [Fin80], [Don82]), built to account for any kind of interpretation pattern, at the cost of several *ad hoc* devices[4] and a critical increase of the size of the lexicon.

Our goal is to design a general model, based on non-specialized lexical information, in order to exhibit the principles of nominal compound interpretation. We want particularly to draw a clear line between two tasks:

- the determination of the possible interpretations of a compound given the syntactic and semantic properties of the constituents,
- the selection of the most probable meaning (mainly resulting from a lexicalization process[5]). This selection involves extralinguistic knowledge.

We implement a model that is meant to carry out the first task, and to provide rules that will help to retrieve further semantic information according to the specificity of a given domain. Semantic interpretation of compounds is a crucial issue in various applications (for example in automatic translation). We intend to use our model to assist the text indexing task, and particularly to provide a semantic information that helps to structure the terminology of a domain.

In the first two sections of this paper, we show that our model for the interpretation of English nominal compounds is mainly focused on a rich description of the semantics of nominals, factorized through the use of a hierarchical classification adopting the WordNet[6] lexical database principles. A detailed semantic characterization of nominal constituents allows us to make predictions about their distributional behaviour within compounds. In the first section, we briefly explain which interpretation principles are used to induce the nature

[4]For example, Finin's model [Fin80] of compound interpretation includes the definition of idiomatic rules.

[5]For example, the meaning settled for the compound *whale boat* is: "a boat which is used to hunt whales". Nevertheless, this compound may also be interpreted as "a boat used to transport whales", as in *cattle boat*. The first meaning has been lexicalized, i.e. established from the common use of the compound. See [Bau79] for the description of the various effects of the lexicalisation process.

[6]WordNet is a trademark of Princeton University. Cf. [MBF+90].

of the predicate that captures the relation between the constituents; in the second we focus on the lexical description of nominal constituents, mainly describing the semantic features that are needed in order to retrieve the missing relation. We then discuss the results of the implementation of this model. In the last section, we suggest how a generic model of the interpretation of nominal compounds may be applied to the text indexing task, provided we make use of these principles to match the specificity of a given corpus. We suggest particularly how corpus-based operations may be used to get information on the semantic characteristics of compounds in specific texts.

1 Interpretation principles

Interpreting nominal compounds consists in retrieving the semantic relation between the constituents, on the grounds of their morphological, syntactic and semantic characteristics. The recovering of the missing predicate may be more or less problematic, according to the clues provided by the compound. While it seems rather straightforward to calculate the relation in the examples 1 and 2 below, the interpretation of the two others seems far less obvious[7]:

1. program generator = predicate: *generate*[8](agent: *generator*, theme: *program*)
2. parsing program = predicate: *parse*(agent: *program*)
3. seasickness pill = predicate: *heal*(instrument: *pill*, theme: *seasickness*)
4. antihistamine pill = predicate: *made of*(result: *pill*, source: *antihistamine*)

In the examples 1 and 2, the relation is easily captured since the verbal predicate is made explicit by the presence of a deverbal noun (i.e. a noun morphologically derived from a verb). On the contrary, in the third example, the

[7]Examples 3 and 4 are adapted from [Bau79]. For a detailed description of the semantic representation that is chosen here, see [FS94].

[8]The underlined constituent is the head of the compound. This device is used to differentiate the representation of compounds such as *parsing program* and *program parsing*.

verbal predicate cannot be directly recovered on the basis of regular morpho-semantic operations: the noun *pill* is only semantically linked to the predicate *heal*. We therefore need more lexical information to decide which underlying relation can be inferred from morphologically simple constituents, and to see how such nouns are bound to verbal predicates. The same problem occurs in the last example, which illustrates the fact that a single noun may be linked to several predicates according to the semantic features of the other constituent. Consequently, different rules are required to recover the missing predicate, according to the characteristics of the constituents that enter the compound:

- The predicate is **explicitly** linked to one of the constituents, on morphological, syntactic and semantic grounds. One of the constituents is a deverbal noun, resulting from the suffixation of a verbal base. The predicate is identified with the verbal root. We differentiate two types of deverbals, adapting Bauer's terminology [Bau79]: a deverbal may refer to the accomplishment or the result of the process referred to by the verb (e.g. *parsing*) or it may occupy the role assigned to one of the arguments of the verb and thus refer to one of the actors of the process (agent, instrument, e.g. *generator*, or patient, e.g. *employee*). In the former case (*action deverbals*), the deverbal inherits the entire argument structure of the verb: *parse*(agent, theme) → *parsing*(agent, theme), and in the latter (*subject deverbals*), it inherits the structure deprived of the argument controlled by the suffix: *parser*(theme).

  When the deverbal noun fills the head of the compound, the modifier (i.e. the leftmost constituent) may satisfy one of the arguments mentioned in the argument structure of the deverbal and fill a thematic role: *sentence parsing* → predicate: *parse*(theme: *sentence*), or it may fill a semantic role, referring to a circumstance of the action (location, time, means, etc.): *hand parsing* → predicate: *parse*(means: *hand*).

  When the deverbal noun is on the left of the compound (i.e. when it occupies the modifier position), it cannot satisfy its object argument within the compound; in this case, the head may only fill a semantic or agentive role: *parsing program* → predicate: *parse*(agent: *program*).

  For a detailed account of these principles, see [Seb93], [Sel82], [Lie83].

- The predicate is **implicitly** linked to one of the constituents, on semantic grounds. Two possibilities arise:

  - some root nouns (i.e. morphologically simple) are typically linked to a verbal predicate, though this relation does not rest on a derivational process. This predicate corresponds to the typical function of the noun's referent. Like subject deverbal nouns, these nouns play a role in the event denoted by the verb they are related to and of which they fill an argument. T.W. Finin [Fin80] gathers these two categories of nouns under the label *role nominals*. As for us, we use this term to refer specifically to non-deverbal nouns alone. For example, the noun *pill* typically refers to the subject argument of the verb *heal*, filling the instrument role. We note this association[9] as follows: *heal*(instrument: *pill*).

    Unlike subject deverbals, role nominals are not provided with an argument structure that may be syntactically satisfied. Nevertheless, within nominal compounds, the argument structure of the underlying verb provides a clue for the distributional capacity of the noun within a compound, so that we may say that these nouns possess semantic arguments (following suggestions made in particular by [Gri91]). For example, the verb *heal* requires a subject and an object argument; since the noun *pill* refers to its first argument, the position which is left empty (the theme) may be occupied by the first constituent of a compound of the form *N pill*, as in *seasickness pill* → predicate: *heal*(instrument: *pill*, theme: *seasickness*).

    This may also be the case - though far less frequently - when the role nominal is the left constituent of the compound: *billiard table* → predicate: *play*(theme: *billiard*, locative: *table*). The noun *billiard* is typically associated to the verb *play*, filling the object argument. The noun *table* designates the circumstance of the action, filling one of the semantic role of the verb *play* (locative).

  - nouns may be bound to verbal predicates through other semantic associations, with respect to different semantic dimensions. For

[9] This semantic association between a noun and a verb is closely related to Melcuk's notion of lexical rules [Mel84].

example, we have seen that the noun *pill* is related to the verb *heal* which refers to its function; but it may enter other kinds of relations which are also at stake in the compounds: as any artifact, this noun may be viewed as playing the object role of a *made of* process, which underlies for example the compound *antihistamine pill* (i.e. a pill made of antihistamine). This predicative information accounts for the composition of the noun's referent, not for its function. Therefore, a noun may be linked to several relational dimensions, or semantic facets (following Pustejovsky's terminology [Pus91]).

In the next step of the analysis, each recovered predicate constrains the kind of nouns that may satisfy its arguments. These restrictions are formulated in the argument structure through selectional features, which specify the semantic class to which the arguments must belong. They help to choose between several predicates in case multiple interpretations come up: in the case of the nominal role *pill*, the verb *heal* expects a theme argument related to a pathology, whereas the object of the verb *made-of* must refer to a substance. Nevertheless, several predicates may be retained if extralinguistic or contextual information are required to decide between several plausible interpretations.

Given the mechanisms we just described, it is obvious that the information needed to recover the semantic relation between the constituents must be rich with respect to the semantics of the nominal constituents. Our aim is to integrate these data without putting too much weight on the lexicon. Consequently, we now examine what kind of lexical information is needed to link predicative information to nominal constituents.

2 Lexical knowledge representation

We first describe the set of semantic classes we use, since semantic typing is a decisive part of our lexical description. We then explain the way predicative information is connected to nominal constituents.

2.1 Semantic typing of the nominal constituents

Semantic typing is a crucial component of our system. First, the interpretation of nominal compounds requires a detailed semantic sorting of the nouns to handle selectional constraints. Traditionally, these semantic features have been expressed within a restricted vocabulary, limited to very general features such as human, concrete, etc. The need for a richer set of features, incorporating pragmatic knowledge, has led to the building of comprehensive taxonomies. A restricted set of features proves to be of no help to handle the semantics of noun-noun association, because, as mentioned before, the semantic class of the constituents is the only clue to choose between several available predicates. Second, a rich semantic classification is needed to factorize predicative information.

In order to have a set of semantic classes at our disposal, we have chosen the WordNet semantic taxonomy [MBF+90] for two main reasons: first, it provides a rich but non-specialized lexical information. Our aim is to avoid *ad hoc* classifications, built purposedly to fit the compound interpretation rules, and to use a domain-independent semantic taxonomy. Second, the hierarchy enables us to cut the semantic tree and to keep the most generic features or to develop some ramifications when we want to adapt it to fit the characteristics of a corpus (see [FS96]). We can also put to use the depth of the hierarchy to express semantic similarities, and relax to some extent the checking of selectional restrictions, as we explain further on.

The WordNet lexical database partitions the lexicon into four categories (nouns, verbs, adjectives, adverbs). As far as nouns are concerned, WordNet organizes the word forms into word meanings. 48000 clusters are created from 57000 nouns. These clusters are related according to the hyponymy relation (= *is-a* link). The hierarchy has several entry points, referring to very general classifications (ENTITY, ACT, STATE, EVENT, LOCATION, etc.). The hyponymy relation generates a hierarchical semantic organization. Consequently, if we choose to cut the tree at a certain level, we can get a set of generic semantic classes. We use indeed only a small part of this hierarchy to define the set of semantic features we need for the interpretation of nominal compound, because many levels are much too specified to categorize nouns in a non-specialized

context, as illustrated on one part of the tree on figure 1. We see on this figure how the classes that convey overdetailed information are discarded. For example, it seems unrelevant to distinguish the alcoholic beverages among the BEVERAGE class.

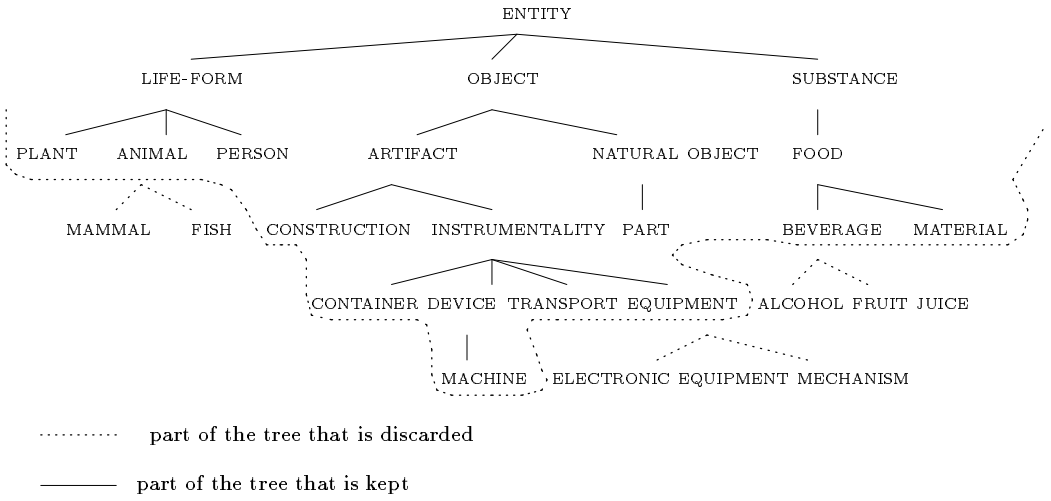

Figure 1: Limiting the depth of the semantic hierarchy

Each noun is then connected to the hierarchy through its semantic label. For example, the word *computer* is typed as MACHINE and is linked to the superordinates: MACHINE → DEVICE → INSTRUMENTALITY → ARTIFACT → OBJECT → ENTITY. Each semantic group in WordNet is defined by a textual gloss, e.g. MACHINE = "any mechanical device that performs or assists in the performance of human tasks".

We don't use all the ressources of the WordNet database. The hierarchy is also organized according to a meronymy relation (i.e.*part-of* relation) which may be very useful since it is one of the standard semantic relations that occur in compounds: e.g. the word *brake* is mentioned as part of a WHEELED VEHICLE, and this information may be used to induce the relationship between the constituents of the compound *car brake*. Yet we have not used this relation so far, because this information is not exhaustive.

Let us see how semantic typing applies to *selectional restrictions*: the arguments of the verbs are semantically labelled, in order to control the way argument positions are satisfied. For example, we have seen that the noun *pill* is linked to two predicates: *heal*(instrument: *pill*) or *made of*(result: *pill*). Since *pill* fills the subject role of the verb *heal*, its object argument is left unsatisfied. In the second case, it is the source argument of the verb *made of* that remains unfilled. In a compound of the form N *pill*, these positions may be occupied by the modifier, provided it meets some semantic constraints. The argument of the verbs may indeed be typed as follows: *heal*(theme: PATHOLOGY) and *made of*(source: SUBSTANCE). It means that the theme argument of the *heal* event must refer to a pathology, and the source argument of the *made of* event must be a substance. These selectional restrictions allow to choose between the two verbs according to the semantic class of the modifier in the compounds where the noun *pill* occupies the head position. Hence the interpretations given above for *antihistamine pill* and *seasickness pill*.

2.2 Associating predicative information to nominal constituents

Two types of predicative information are related to nouns:

- a functional predicative information, which bears on the function assigned to the entity which the noun refers to (e.g. *pill* has to do with a *healing* process). This information is common to subject deverbals and role nominals.
- a non-functional predicative information, which conveys other types of relations between nouns (e.g. one way to characterize the object *pill* is to indicate what it is made of). This information pertains to all types of nouns, including deverbals.

We first give details of this distinction, which has a lot to do with Pustejovsky's notion of semantic facets [Pus91]. Second, we explain how this predicative information is part of the lexical definition of the nouns. We distinguish two cases:

- The predicative information appropriate for a noun cannot be generalized to a whole class of nouns. It appears directly in the lexical entry of the noun. For example, the predicate *wash* is typically associated to the noun *soap*, via its intrument role. This information cannot be associated to a whole class of nouns (unless we define a class of "cleanser"[10]) and is thus specific to the lexical entry of the noun *soap*.
- The predicative information may be generalized to a whole class of nouns. It is moved up to the level of the class itself. For example, the noun *box* shares the information "instrument of a *contain* process" with any container. If the class *container* is available (and it is in our model) then this link to the predicate *contain* may be expressed at the level of the class.

2.2.1 Different types of implicit predicative information

Our model relates strongly to Pustejovsky's views on the semantics of nominals, whose model allows to capture the various semantic facets of the nouns through reference to the relations in which the nouns can enter. One crucial component of his generative lexicon is the *qualia structure* [Pus91], which specifies nominal facets through the use of subtyping, showing the relations associated to nouns. These subtypes are CONSTITUTIVE ("the relation between an object and its constituent parts"), FORMAL ("that which distinguishes the word within a larger domain"), TELIC ("its purpose and function"), AGENTIVE ("factors involved in its origins"). Some features of this structure may be illustrated on the word *pill*, as shown on figure 2.

This representation states that PILL is a MEDICINE, that its component elements are defined as SUBSTANCE and that its purpose is an activity of healing. The formal parameters may be typed (e.g. z: PERSON). Pustejovsky's model is used particularly to induce collocational preferences. We claim that this structured representation, that account for the semantic behaviour of the nouns, is particularly well adapted to handle the semantics of nominal compounds, and that these semantic facets must be recovered to calculate the possible subjacent relations within nominal compounds.

[10]The class CLEANSING AGENT, CLEANSER is available in the WordNet hierarchy but we discard it because it bears on a too specific semantic information.

$$\begin{bmatrix} \text{pill(x,y)} \\ \text{CONSTITUTIVE= SUBSTANCE(y)} \\ \text{FORMAL = MEDICINE(x)} \\ \text{TELIC = HEAL(x,z)} \end{bmatrix}$$

Figure 2: Semantic facets of the word *pill*

In what follows, we see how these semantic relations are represented in our model, either directly at the level of lexical entries, or through the semantic description of the classes in the hierarchy.

2.2.2 Predicative information appearing in the lexical entries

Two types of nouns may be associated to specific predicative information, which cannot be generalized at the level of semantic classes: deverbals, and role nominals.

- The lexical information concerning the **deverbals** is recovered from the verb and the suffix, unless some irregularity occurs, in order to limit redundancy. Action deverbals are mainly *-ing*, *-ation* or *-ment* nominalizations (e.g. *parsing*, *reformation*, *disarmament*). They inherit the entire argument structure of the verb, and are typed as ACTION.

  parse + *ing* =

  parsing: agent(PERSON), theme(COMMUNICATION), sem(ACTION)

  The action deverbal *parsing* inherits the argument structure of the verb *parse* (by default, the agent is specified as a PERSON. The WordNet COMMUNICATION class, which labels the theme, includes in particular examples of linguistic communication, such as *sentence*), and gets its semantic class ACTION from the *-ing* suffix.

  Subject deverbals result mainly from *-er* suffixations, though *-or* and *-ant* suffixations are also available, but less productive (e.g. *parser, generator, attendant*). The deverbal refers either to a human or to an instrument, and is thus coded twice, as PERSON and as INSTRUMENTALITY.

parse + *er* =

parser-1: theme(COMMUNICATION), sem(PERSON),

parser-2: theme(COMMUNICATION), sem(INSTRUMENTALITY)

The subject deverbal *parser* inherits the argument structure of the verb *parse* minus the subject argument which is controlled by the suffix. Two entries are generated, each corresponding to one sense of the *-er* suffix.

When a deverbal shows some irregularity with respect to this standard derivational process, it is listed in the lexicon and all these information are mentioned extensively in its lexical entry. It is the case in particular when the deverbal selects only some of the senses of the original verb, e.g. *coverage*, or when a subject deverbal is generated from a suffixation usually used to derive action deverbals, e.g. *government*.

- The lexical entry of **role nominals** includes both the predicate typically associated to the noun and the role it refers to. Role nominals are mostly:
  - nouns designating the instrument used to perform an activity or a process
    e.g. *steer*(instrument: *rudder*),
  - nouns designating the agent which performs an activity or a process
    e.g. *travel*(agent: *tourist*),
  - nouns designating the place where an activity or a process is performed
    e.g. *gamble*(locatif: *casino*),
  - nouns designating the object of an activity or a process
    e.g. *wear*(theme: *shirt*).

  The lexical entry also includes the argument positions which are not occupied and may be satisfied by the other constituent of the compound: e.g. *cut*(instrument: *knife*, theme: OBJECT).

The predicative information may also be characteristic of a whole class of nouns:

2.2.3 Predicative information moved up to the level of semantic classes

When possible, semantic information is generalized to be captured at the level of semantic classes. In that case, the noun inherits the relational properties of its class. Consequently, we use the WordNet database as a starting point: predicative information is then attached to the semantic nodes of the hierarchy, generalizing semantic facets at the level of semantic classes.

We distinguish between functional and non-functional predicative information:

- semantic classes may gather sets of role nominals, i.e. nouns that share the same purpose, characterized through the same predicate. For example, the CONTAINER class is made up of nouns referring to objects whose function is to *contain* something (*bin*, *bottle*, *envelope*, etc.). The commun feature shared by the members of such classes is thus the telic role in Pustejovsky's terms. Following the partition we previously present, we may distinguish:

  - classes of role nominals designating the instrument used to perform an activity or a process. Any instrument is associated to a functional predicate and enters the generic relation: INSTRUMENTALITY = *for*(instrument: INSTRUMENTALITY, theme: _[11]). This predicate is specialized according to the type of action that the instrument helps to perform, e.g. VEHICLE = *transport*(instrument: VEHICLE, theme: ENTITY)
  - classes of nouns designating the agent who performs an activity or a process. Such classes exist in WordNet (e.g. SELLER), but are too specific to be part of our hierarchy, which does not go below the node PERSON.
  - classes of nouns designating the place where an activity or a process is performed: WORKPLACE = *work*(agent: PERSON, locative: WORKPLACE) (e.g. *laboratory*),

[11] This symbol indicates that the argument is semantically unrestricted.

– classes of nouns designating the object of an activity or a process: GOODS = *commercialize*(agent: PERSON, theme: GOODS) (e.g. *clothes*).

Such classes are thus related to a predicative relation, as illustrated on figure 3.

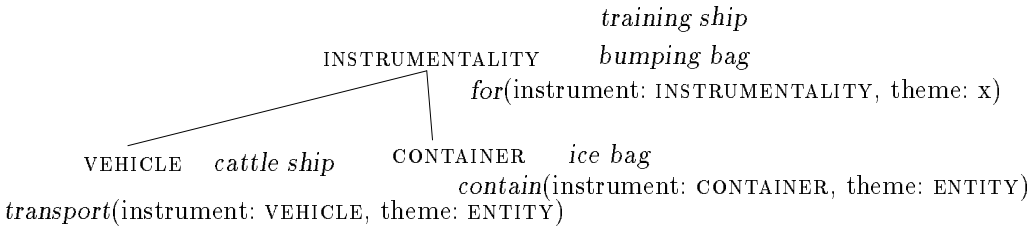

Figure 3: Relating role nominal classes to predicates

- semantic classes may gather nouns that are characterized by a non telic semantic link to other nouns. For example, *part-whole* (characteristic of the class PART, e.g. *balance arm*), *whole-component* (class ARTIFACT, e.g. *vegetable oil*), *object-attribute* (class ATTRIBUTE, e.g. *oil prices*) relations. WordNet classification helps to define classes of nouns that are typically involved in this kind of semantic patterns. One may distinguish two classes of such relations: constitutive and characterizing relations[12].

  – constitutive relation:

  ARTIFACT = *made-of*(object: ARTIFACT, source: SUBSTANCE) (e.g. *antihistamine pill*)

  INSTRUMENTALITY = *made-of*(object: ARTIFACT, part: DEVICE) (e.g. *saw grinder*)

[12]In what follows, the representation of the meaning may slightly differ from what we have seen so far. It is indeed very unsatisfactory to label the arguments of this kind of predicates in terms of thematic roles, hence the use of pragmatic labels, which convey a more specific and relevant information. In some examples, the ordering of the arguments may be sufficient to characterize the relation; in that case, we use variables instead of labels.

SUBSTANCE = *constitute*(source: SUBSTANCE, object:ARTIFACT) (e.g. *gluten bread*)

SUBSTANCE = *constitute*(source: SUBSTANCE, object:ARTIFACT) (e.g. *pastry flour*)

PART = *constitute*(part: PART, object:ENTITY)(e.g. *eagle feather*)

– characterizing relation:

REPRESENTATION = *represent*(x: REPRESENTATION, y: _) (e.g. *sentence pattern, peace emblem*)

KNOWLEDGE = *about*(x: KNOWLEDGE, y: _) (e.g. *education system, blood test*)

ATTRIBUTE = *characterize*(x: ATTRIBUTE, y: _) (e.g. *desk height*)

This list is not closed. These relations are based on the predicative information that is found in the textual gloss associated to each class in WordNet:

SUBSTANCE = "the tangible stuff *of which an object consists*"

PART = "any part *of an animal or plant* such as an organ or extremity"

The treatment of abstract nouns poses a real problem, and is generally neglected by semantic models of nominals. It is particularly problematic to define their semantics in term of facets according to the semantic grid defined by Pustejovsky in the qualia structure, which mainly deals with nouns referring to what we commonly may define as "objects". The telic role, denoting the function for which an object as been conceived, is hardly transposable to the set of abstract nouns. Consequently, we have mostly based the semantic description of these nouns on the information contained in the textual gloss of WordNet classes, in order to link possible predicates to the nodes which dominate abstract nouns in the hierarchy (i.e. mainly the nodes subsumed by the categories ABSTRACTION and PSYCHOLOGICAL FEATURE). For example, the word *height* is an instance of the class ATTRIBUTE, defined in WordNet as "an abstraction belonging to or characteristic of an entity". Each member of this class may thus potentially collocate with nouns denoting the entity that

is characterized, hence the interpretation: *desk height* → predicate: *characterize*(attribute: *height*, entity: *desk*).

A noun may therefore be attached to several predicates, on account of its inclusion in several semantic classes of the hierarchy, as illustrated in figure 4.

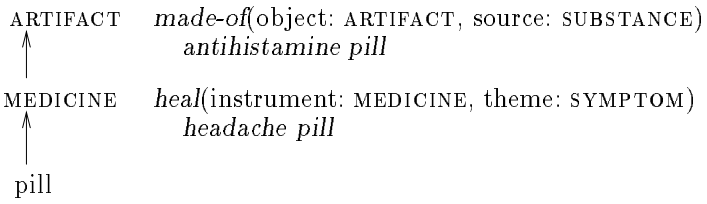

Figure 4: The word *pill* attached to the lexical hierarchy

This association of non-functional predicates to nouns applies also to deverbals, which are semantically typed according to the nature of their suffixation. For example, the noun *generator*, which is typed as INSTRUMENTALITY, inherits the relation *made-of* in addition to its morphological link to the predicate *generate*. This facet is illustrated in the compounds *program generator* and *valve generator* (predicate: *made-of*(object: *generator*, part: *valve*)).

We have shown how we enrich the lexical description of the nouns to deal with the semantics of nominal compounds, without dangerously increasing the length of the lexicon. We now turn to the description of the first results obtained by a program which experiments this model.

3 Implementation of the model and tests

Our modeling of the semantics of nominal compounds has been implemented and tested on a list of 100 sequences randomly picked up from a large corpus of 9000 NN compounds built up by R. Sproat [Spr94] from texts related to various domains. These sequences are deconnected from their original context, so that the point is not to automatically calculate their precise meaning as their appeared in the text from which they have been extracted, but to generate any meaning that could be predicted considering the constituents.

We first study examples of interpretations generated by our program before discussing the problems encountered. For each example, we briefly examine the following points:

- does the program generate several interpretations?
- what characteristics of the constituents trigger the interpretation(s)?
- are they correct?
- is there any other interpretation missing?

3.1 Description of the results

Let us first enumerate the interpretation rules that we refer to, in order to facilitate the explanation of the examples.

1. <N> + <V-act> → predicate: V(role: N)

   The compound is headed by an action deverbal; the semantic relation is given by the verbal base; the modifier satisfies one of its arguments (thematic or semantic).

2. <N> + <V-suj> → predicate: V(role-subj: N, role: N)

   The head is a subject deverbal; it controls one of the arguments of the verbal base.

3. <V-act> + <N> → predicate: V(role-sem: N)

   The modifier is an action deverbal; the head constituent occupies one of its semantic arguments.

4. <N> + <Nrole> → predicate: V-role(role: Nrole, role: N)

   The predicate is linked to the head constituent, which is a role nominal.

   The last rules account for non-functional semantic relations, linked to root nouns according to their semantic class.

5. <N> + <COMMUNICATION> → predicate: *about*(communication: COMMUNICATION, theme: N)

6. <SUBSTANCE> + <ARTIFACT> → predicate: *made-of*(object: ARTIFACT, source: SUBSTANCE)
7. <SUBSTANCE> + <SUBSTANCE> → predicate: *from*(substance: SUBSTANCE, source: SUBSTANCE)
8. <N> + <ATTRIBUTE> → predicate: *characterize*(x: ATTRIBUTE, y: N)
9. <N> + <DEVICE> → predicate: *for*(device: DEVICE, goal: N)
10. <N> + <CONTAINER> → predicate: *contain*(instrument: CONTAINER, theme: N)
11. <N> + <COVER> → predicate: *cover*(instrument: COVER, theme: N)
12. <N> + <KNOWLEDGE> → predicate: *about*(knowledge: KNOWLEDGE, theme: N)
13. <INSTRUMENTALITY> + <ACTIVITY> → predicate: *perform*(instrument: INSTRUMENTALITY, theme: ACTIVITY)

We now show examples for each category of compounds, that is compounds exhibiting an explicit or an implicit predicative relation.

- <N> + <action-deverbal>

  1. *earthquake coverage* →
     The program generates two interpretations:
     a. predicate: *cover*(theme: *earthquake*)
     b. predicat: *about*(communication: *coverage*, theme: *earthquake*)
     Following WordNet definitions, the deverbal *coverage* has two entries in our lexicon, one corresponding to the meaning *reportage* (class COMMUNICATION) and the other to the meaning *insurance* (class POSSESSION). Although it is an action deverbal, we have chosen to list it in the lexicon because only 2 out of the 15 senses of the verbal base are retained by the derivation process. The noun *earthquake* is defined as a NATURAL PHENOMENON.

Both interpretations correspond to the first meaning of *coverage* (rule 1). The second interpretation is due to the fact that *coverage* is typed as COMMUNICATION, thus inheriting also the relation which is characteristic for this class (rule 5). These two interpretations are very similar, the second one being more precise (the predicate *cover* is one instance of the generic predicate *about*).

Another interpretation would be expected, on the basis of the second meaning of *coverage*, i.e. "reimbursement in the case of loss". *Earthquake coverage* could mean "reimbursement in the case of loss, after an earthquake". We see that this relation is more far-fetched and demands a very sophisticated extralinguistic reasoning about the possible relationships between the two nouns.

2. *weekend slayings* → predicate: *slay*(time: *weekend*)

   Only one sense is generated, since the selectional restrictions on the arguments of the verb enable us to decide that the modifier occupies a semantic (temporal) role: the verb *slay* selects an animate object, which is incompatible with the semantics of the modifier *weekend* (rule 1). Selectional constraints enable us to discard unplausible interpretations (such as predicate: *slay*(theme: *weekend*)).

- <N> + <subject-deverbal>

  wood checker →

  (a) predicate: *check*(agent: *checker*, theme: *wood*)

  (b) predicate: *check*(instrument: *checker*, theme: *wood*)

  (c) predicate: *made-of*(object: *checker*, source: *wood*)

  The features of the subject deverbal *checker* are automatically inferred from the lexical entries of the verbal base *check* and the suffix *-er*. Two instances of the deverbal are generated, one with the feature PERSON, the other with the feature INSTRUMENTALITY. Interpretations (a) and (b) are instances of rule 2 and involve each entry of the deverbal. Besides, since the deverbal is labelled INSTRUMENTALITY, rule 6 applies and generates the third interpretation.

Multi-generation is a normal side-effect of our model. If we wish to limit this phenomenon, we need to add biases to our set of rules. In particular, we may introduce a notion of precedence, and state for example that when the program generates an interpretation related to the verbal base of the deverbal, then it must be produced preferentially to a rule involving the semantic class of the deverbal.

This kind of heuristics may be inferred from statistical observations.

- <action deverbal> + <N>

  backup weapon →

  predicate: *backup* (instrument: *weapon*)

  predicate: *perform*(instrument: *weapon*, theme: *backup*)

  The first interpretation comes from rule 3, the second from rule 13 since the label ACTIVITY is assigned to the action deverbal *backup*. The two interpretations are similar apart from the level of precision.

- <N> + <role nominal>

  1. *bullet casings* → predicate: *cover*(instrument: *casing*, theme: *bullet*)

     The word *casing* is listed twice in the lexicon, first as a COVER ("the outer covering or housing of something") and second as a CONSTRUCTION ("the enclosing frame around a door or window opening"). No relation is available between the noun *bullet* and the second meaning of the word *casing*, which is thus discarded. The interpretation of the compound *bullet casings* is thus based on the definition of *casing* as an element of the class COVER, which gathers role nominals filling the instrument role of the predicate *cover* (rule 11).

     This example illustrates the fact that the association of two nouns may contribute to disambiguate them, in case they are polysemic.

  2. *liquor bottles* →

     predicate: *hold*(instrument: *bottle*, theme: *liquor*)

     predicate: *made-of*(object: *bottle*, substance: *liquor*)

The first interpretation is generated by the rule 10, on account of the relation between the class CONTAINER and the predicate *hold*. Yet, the noun is also linked to the upper node ARTIFACT, which may enter a constitutive relation, provided the modifier is a SUBSTANCE (rule 6). This double connection is illustrated on figure 5. A second interpretation is thus generated, corresponding to the paraphrasis "a bottle which is made up of liquor" (on the same pattern as the compound *paper bag*). This interpretation is surely mistaken, on account of extralinguistic knowledge (i.e. liquor cannot be a material), but this kind of overgeneration is unavoidable and even desirable if we want to ensure the generality of the mechanisms we apply.

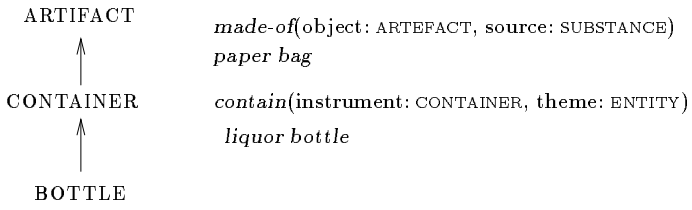

Figure 5: The word *bottle* attached to the lexical hierarchy

3. *rudder controls* → predicate: *for*(object: *controls*, destination: *rudder*)
   The noun *controls* is a role nominal associated to the predicate *regulate*: *regulate*(instrument: *controls*, theme: ACTIVITY), according to the WordNet definition: "a mechanism that *regulates* the operation of a machine". The program fails to generate the expected interpretation (i.e. predicate: *regulate*(instrument: *controls*, theme: *rudder*) since the modifier *rudder* does not match the selectional constraint associated to the theme. Rule 4 does not apply, but a less specific interpretation is generated, through the rule 9, which is triggered on account of the semantic class of the head.

- <N> + <N>

  In the last three examples, the semantic class of the head is the clue which triggers the interpretation:

1. *desk height* → predicate: *characterize*(attribute: *height*, entity: *desk*)

   The noun *height* is typed as ATTRIBUTE. This class is linked to a characterizing relation which applies here. Only one interpretation is generated, in accordance with the non-ambiguity of the compound.

2. *campaign motto* → predicate: *about*(knowledge: *motto*, theme: *campaign*)

   The noun *motto* is labelled as KNOWLEDGE, which defines a constitutive predicate *about*, referring to the content of the message that is communicated. The semantic representation in this case is very general, illustrating the different levels of information available on the nominal constituents. Three identical representations are generated, corresponding to the three entries of the noun *campaign*: SOCIAL EVENT = "a race between candidates for elective office", SOCIAL ACTIVITY = "military campaign", MOVEMENT = "hunting expedition, safari".

   To be more accurate, the interpretation should be based on a more detailed lexical definition of the word *motto*, which refers to the political domain (first meaning of *motto*), but our model does not allow for the integration of such data.

3. *wood ashes* → predicate: *from*(substance: *ash*, origin: *wood*)

   This is an instance of rule 7.

We now go further into the description of some of the problems encountered in our general framework and brought out by this test.

3.2 Evaluation of the results

It is not always possible to precisely evaluate the accuracy of the resulting representation. An English speaker may almost always conceive a new meaning from a given compound. For example, if the expected meaning of *television program* is "a program *on* the television", we may still suggest another meaning, such as "a program *about* television". The second meaning belongs to the semantic possibilities of the head noun *program* (as in *art program*). Therefore, we may only judge the plausibility, not the correctness, of the answers,

which form an open set. This program is thus a preliminary but necessary step towards our aim, which is to adjust the interpretation to the specificity of the corpus, on the basis of a general, portable semantic modeling.

Nevertheless we are able to globally estimate the results of our program: out of 100 compounds, the program fails to produce any answer on 29 compounds. If we eliminate inaccurate sequences (*Crude prices, Toot Toot, studies series*) and lexicalized compounds (*spill laws, hour dresser, business names*), we end up with 23 missing interpretations, mostly because:

- no generalization has been made about the semantic class of the constituents, e.g. *loss ratio* (class of the head = RELATION), so that the underlying relationship does not fit any of the standard semantic relations that are registered. In particular, some of the abstract semantic classes gather heterogeneous types of nouns, and it is very difficult to define their collocational properties. The semantics of the compound may anyway be difficult to account for in terms of a predicative relation (e.g. *marathon tour*, which illustrate the *subclass* relationship.).

- a constituent has undergone a semantic shift not registered in the lexicon. This problem results in the violation of selectional restrictions. For example the noun *rail* refers to the class SOCIAL GROUP ("commercial organization responsible for operating a railway system"). In *rail travel*, it is therefore interpreted as the agent of the action of travelling, whereas it should be understood metonymically as a transport, hence as a circumstance of the action. The interpretation of compounds faces the traditional problems of nominal polysemy and must take into account these regular alternations. For example, the word *court* may refer just as well to the assembly which conducts judicial business or to the room where it is conducted. This classic group/location alternation occurs in other words of the corpus such as *circus*, *laboratory*, *stadium*, and must be handled through specific lexical rules (see [CB95] on this issue). Other semantic shifts are more difficult to predict, such as the one which affects the word *weapon* in the compound *weapon workers* (the product of an industry refers to the industry itself).

In what follows, we develop these two problems, and we examine the issue of multi-generation that has been raised in the previous subsection.

3.2.1 Unpredicted semantic patterns

Our aim is to base our model on a very general semantic knowledge. In many cases, the resulting interpretation is thus under-specified. For example, the relation *about* will be assigned to all compounds in which the head belongs to the class KNOWLEDGE, such as *disarmament program*, *casino issue*, *circus tradition*, *research data*. Some compounds matching this pattern would require a more specific interpretation, e.g. *court matters* (*matters* dealt with by the *court*). As a consequence, a too general model fails to record the specific behaviours of certain classes of nouns. For example, if we consider the pattern N + COMMUNICATION, in which the class COMMUNICATION denotes "something that is communicated between people or group of people", we interpret the modifier as the theme of the communication process:

artillery contract → predicate: *about*(communication: *contract*, theme: *artillery*)

earthquake coverage →
predicate: *cover*(theme: *earthquake*)
predicate: *about*(communication: *coverage*, theme: *earthquake*)

This leads to an erroneous interpretation for the compounds *television program* and *cable program*, since the most plausible interpretation would label the modifier as the source of the communication, not as the theme. This information is available if we integrate the class BROADCAST ("message that is *transmitted by* radio or television") in the hierarchy. As a consequence, specific patterns (such as <TRANSMISSION> + <BROADCAST>) must be identified to refine our model. We propose that these patterns may be acquired from texts, through statistical methods. This hypothesis is discussed in the last section.

3.2.2 Violation of selectional restrictions

The head noun controls the modifier via the selectional features which restrict the set of items that may occupy the positions specified in the argument structure. The idea is to find a balance between an excessively restrictive view of the selectional restrictions (e.g. what may be found as object of the verb *eat* is restricted to the nouns belonging to the class FOOD) and a laxist view which leads to deny any controlling capability to the selectional constraints (e.g. any noun which falls into the scope of a verb gain the semantic features which are

required by any noun in such a position, in order to consider unexpected associations such as *opium eater*, *spider eater*, etc.). The latter view is unrelevant when the aim is to use the selectional restrictions as a means to disambiguate a construction. It is precisely the point here: when a constituent is linked to several predicates, the semantic features of the other constituent are the main clue to decide between theses alternatives. Besides, selectional control is required to decide between thematic and semantic relations (*cattle feeding* vs *rod feeding*) or between two different thematic roles (e.g. *cattle feeding* vs *corn feeding*)[13].

This problem is encountered on two examples: *art collection*, *juice exports*. The program fails to give any interpretation because the semantics of the modifier does not match the constraints on the object argument of the verb: in *juice exports*, the argument is typed as GOODS and the modifier as FOOD.

As a consequence, we need to relax the control expressed via the selectional features (to take into account what Pustejovsky calls "licensed violation" of selectional restrictions). The difficulty is to determine to what extent. As Resniks [Res93] points out, the evaluation of the semantic similarity between nouns is an issue. If we use the semantic hierarchy, we may relax the unification process and consider that the modifier may satisfy the argument position if:

1. the modifier belongs to the class mentioned by the selectional features.
2. the modifier belongs to a more specific class (e.g. in *pie eater*, *pie* is linked in WordNet to the class BAKED GOODS, which is a daughter of the class FOOD),
3. the modifier is semantically similar to the set of nouns required by the predicate, i.e. it belongs to one of the class subsuming the class mentioned by the selectional restrictions. For example, if the verb *export* requires a theme element of the class GOODS, and that it is associated to the word *juice* in the compound *juice export*, then the modifier will be considered

[13]We have noticed that when the predicate is implicit, the range of the nouns that can be found at the modifier position filling the object role is more restrictive. For example, when the role nominal *spoon* is the head of a compound, the modifier strictly belongs to the class FOOD in our corpus: *salad spoon*, *cream-soup spoon*, etc. These compounds are based on a shared stereotypic knowledge which restricts the semantic diversity of the constituents.

as the object of the verb, since the class GOODS and the class FOOD share a common upper-class OBJECT.

This is of course a first approximation. We have to define the limit of this process, and particularly cases of incompatibility (for example, OBJECTS and PERSON are linked to a common upper-node ENTITY, but it is undesirable to allow the unification process to merge these two classes, which generally refer to different kinds of roles). To go further, corpus-based deductions are required, to define more accurately semantic similarity (cf. [Res93] on this issue).

3.2.3 Multi-generation

Let us observe the result of the interpretation when the constituents are polysemic. For example, if we take the compound *court supervision*, we see that the noun *court* may refer to a SOCIAL-GROUP (court-1 = "an assembly to conduct judicial business") or a CONSTRUCTION (court-2 = "a room in which a law court sits"). In that case, two types of polysemy occur: the polysemy of the constituents and the polysemy of the relation.

court supervision →
predicate:*supervise*(theme: *court-1*)
predicate: *supervise*(agent: *court-1*)
predicate: *supervise*(theme: *court-2*)
predicate: *supervise*(locative: *court-2*)

Three interpretations come up, corresponding to three possible argument links: the supervision OF the court (theme), the supervision BY the court (agent) and the supervision IN the court (locative). Consequently, 2*3 representations are likely to be generated. The association of the two words discards two possibilities (i.e. court-1, referring to an animate entity, cannot satisfy the locative role, and court-2, referring to an object, can not satisfy the agentive role) but the remaining combinations show how polysemy can rapidly increase the number of the interpretations that are produced.

Yet this is a desirable result: semantic information is unsufficient to choose between these interpretations. In that case, examples of similar but non-ambiguous compounds would help to eliminate some of the possibilities or to decide in favour of one of them. In the corpus made up by R. Sproat, we find two sequences: *government supervision*, *community supervision*, in which the

modifier refers unambiguously to a SOCIAL GROUP; this observation could be used to favor the first two interpretations.

Thus, information specific to the corpus may help to guide the interpretation process. This is the hypothesis that we discuss in the last section.

4 Future works: applying general interpretative principles to the treatment of specific corpora

In what follows, we make suggestions on how corpus-based operations may help to refine the interpretation process, through the acquisition of specific information that are relevant for the interpretation of the compounds of a given corpus. These hypotheses are being currently implemented and tested, but they are based for the moment on observations on a small corpus.

We finally illustrate how the interpretation of nominal compounds may contribute to enrich text indexing.

4.1 Refining the model: extracting semantic information on the compounds of specific texts

First we show briefly to what extent compounds in domain-specific texts match the general principles we have described so far, and to what extent they differ (less ambiguous constituents, compounds displaying specific relationships). Second we describe similar works dealing with lexical semantics and using corpus-based methods. Finally we make suggestions on how statistical and corpus-based observations may contribute to the interpretation of nominal compounds in the general framework that we have introduced.

4.1.1 Compounds in domain-specific texts

If we list the compounds of a technical text such as a handbook in computer science[14], we may confirm the generality of the principles we have described. In particular, the compounds may be classified according to the patterns we have presented:

- N + deverbal: *file loading, programmer request, file compiler*
- deverbal + N: *operating systems*
- N + role nominal: *system interface*. The noun *interface* is typically linked to the predicate *connect*: *connect*(instrument: *interface*, theme: ENTITY)
- N + N
  - constitutive relation: *error message* (predicate: *about*(communication: *message*, theme: *error*), *directory component* (*part-of* relation).
  - characterizing relation: *file names* (predicate: *characterize*(x: *name*, y: *file*).

What we must take specifically into account is:

- lexical information: the interpretation task is made easier since it deals with less ambiguous nominal constituents. For example, WordNet lists four senses for the noun *file* and ten for the noun *set*, which are monosemic in this corpus.
- collocational information: specific semantic patterns are found in the corpus. For example, the compounds *UNIX file system* or *Lisp objects* are instances of a very common pattern in the text.

[14]The corpus we use is the online version of Guy L. Steele "Common Lisp the Language", Second Edition.

4.1.2 Retrieving semantic information from corpora: related works

Statistical techniques are frequently used to retrieve lexical information from texts, such as semantic typing of nouns and adjectives [AB95] [AB96], subcategorization frames [Res93] or taxonomic relationships [AB96] [PAB93]. The idea is that a pre-established lexical representation of the words of a text is not always available or that it fails anyway to grasp the specificity of a given text or set of texts belonging to the same domain. It is especially the case when one wants to retrieve information from large corpora of technical texts: lexical or conceptual information about the content of these texts are not always available.

We develop here two examples, namely semantic relations retrieval and semantic typing, which illustrate these techniques

Pustejovsky et al. [PAB93] show how statistical techniques, such as mutual information measures (which is used to assess the degree of association between words), can help to discover semantic relations between words. They mention the example of nominal compounds based on a taxonomic relation, i.e. a subclass relationship, such as *Unix operating system* and *C language*. This type of semantic relation is indeed very difficult to infer, since it is entirely based on pragmatic knowledge. This information is to be found at the bottom of the WordNet lexical tree[15], reflecting encyclopedic knowledge which is ever increasing and which varies according to the domain. Since one can not list all the members of such sets as OPERATING SYSTEMS or LANGUAGES, one may wish to infer the membership relation from texts. The idea is to deduce this information from the context, seeking for word associations displaying an explicit and non ambiguous relationship. Pustejovky et al. propose to use information regarding verb subcategorization: if a noun N1 is a member of a class denoted by the noun N2 (e.g. C is a member of the class LANGUAGE), then the two nouns are likely to appear in the scope of the same verbs. The similarity score between the two nouns is computed on the basis of the mutual information value between those nouns and the verbs with which they

[15]For example: C, COBOL, BASIC → PROGRAMMING LANGUAGE → ARTIFICIAL LANGUAGE → LANGUAGE → COMMUNICATION → SOCIAL RELATION → RELATION → ABSTRACTION

co-occur. Corpus-based operations are thus used to validate or invalidate some semantic relations.

Statistical methods may also help to semi-automatically retrieve the semantic typing which characterizes the words - and particularly the nouns or nominal terms - of a given text. The set of semantic classes that are needed to type the nouns of a text depends highly on the set of objects that characterizes the domain. For example, in the domain of electric networks, Assadi and Bourigault [AB95] show that semantic classes such as CONNECTION or PHYSICAL MEASURES are needed to characterize the nouns or nominal groups that appear in the corpus. Statistical methods of data analysis are used to outline these semantic sets, the results being communicated to a terminologist in order to be validated and labelled or unvalidated.

In what follows, we suggest how similar techniques may be applied in addition to our rule-based model, for the interpretation of nominal compounds in specific texts.

4.1.3 Acquiring semantic patterns

Statistical information has been taken into account by several psycho-linguists such as P. Downing [Dow77] and M.E. Ryder [Ryd94]. Their purpose is to use statistical knowledge to interpret novel pairings. Ryder claims that a set of semantic rules is not sufficient to deal with the productivity of the compounding process, since the creation of new compounds involves extralinguistic knowledge and cognitive strategies. According to her, "the predictability is probabilistic", and she shows that the creation and interpretation of novel compounds is based on knowledge about productive semantic patterns. For example, she lists highly frequent templates such as:

1. ANIMAL1 + ANIMAL2 = ANIMAL2 *resembling* ANIMAL1
2. BODY PART + NON-CLOTHING = NON-CLOTHING *that is operated by/used on* BODY PART (*foot pedal*, *hand soap*, *neck brace*)
3. X + PRODUCT = PRODUCT *used on* X (*pet shampoo*, *laundry detergent*)

The first template cannot be integrated in our model in terms of semantic rules: the predicate *resemble* is not a characteristic relation of the ANIMAL class (the definition of the class - a *living organism characterized by voluntary movement* - only provides the predicate *live*, which is involved in the compounds *beach bird* or *tree cobra*). The second and third patterns give more general information (the compound *pet shampoo* would be interpreted in our model under the condition that *shampoo* is coded in the lexicon as a role nominal: *wash*(instrument: *shampoo*, theme: ENTITY)).

Downing's proposal is very similar: her experiments indicate that the way English speakers tend to create or interpret novel compounds suggests that certain relationships are preferred, depending on the semantic class of the head constituents. She shows for example that animals are characterized first by their appearance (*giraffe bird*) and second by their habitat (*roadside flowers*).

This kind of information is not integrated in our general model of the interpretation of compounds because:

- the compound relation cannot be retrieved on the basis of semantic knowledge on the constituents: the fact that animals are characterized mostly through their resemblance to other animals or objects is part of extra-linguistic knowledge.
- statistical results depend highly on the corpus that has been used to compute them.

Exhibiting semantic patterns in the texts is thus a way to automatically learn these patterns, using for example mutual information procedures. This would also give information about which facet of the noun is mostly used in the compounds of a given corpus (e.g. *made-of* relations, *part-of* relations or functional relations) according to the type of modifiers with which it collocates most of the time.

4.1.4 Searching for contextual information

Contextual information may help to identify the relation between the constituents of the compounds. It may first help to disambiguate a construction. For

example, the interpretation of *compiler processing* will generate the two possibilities: predicate: *process*(agent: *compiler*) and predicate: *process*(result: *compiler*). A look at the surrounding context provides a reliable clue to choose the first interpretation, since the compound is modified by a nominal group which realizes the object argument of the process predicate: *compiler processing of a call*[16]. Second, contextual information may help to identify the missing relation by looking elsewhere in the text to see if the constituents of the compound are involved in another kind of linguistic construction, where their semantic relation would be explicit. Given a compound N1 N2, we may look for strings in which the couple (N1, N2) occurs in a different form. For example, the context may express the relation by the insertion of a preposition:

pathname components: (pathname, components) = "a pathname *with* the same components"

file operations: (operations,file) = "operations *on* the file"

dribble operation = (dribble, operation) = "operation *of* dribble"

or it may provide the missing verbal predicate:

compiler warnings: (compiler,warning) = "it is reasonable for the compiler to *emit* a warning"

file system interface: (file system, interface) = "a standard interface for *dealing with* such a file system"

These are suggestions which have not been explored yet. An observation of the context should be precisely guided to determine the kind of information that is likely to be reliable (prepositional links expliciting constitutive roles, verbal links expliciting telic roles, etc.). One should particularly take into account, using statistical measures, the fact that the context may of course provide a predicative information which is not the one that is involved in the compound (for example for the compound *compiler warnings*: "warnings *deferred* by the compiler").

[16]Note that this construction is forbidden within the generative framework. Grimshaw [Gri91] claims that "compounding of an external argument is impossible when the predicate takes an internal argument in addition". Hence the ungrammaticality of her example: **novices arranging of flowers*, which displays exactly the same construction as the one we mention. This suggests how difficult it is to match pre-defined rules and corpus-based inferences.

We now suggest how the semantics of nominal compounds may be taken into account in the text indexing task.

4.2 Suggestions to enrich text indexing by taking into account the semantics of nominal compounds

The general model we propose may be of great use to organize the nominal compounds of a text in semantic paradigms in order to enrich the description of the objects denoted and eventually help the terminologist to go through the content of a text.

One way to structure the terms of a text is to subdivide the list of terms according to the semantic class of the modifier, exhibiting the types of associations in which the head noun may enter. For example, figure 6 shows how a list of NN compounds found in the corpus of R. Sproat may be structured.

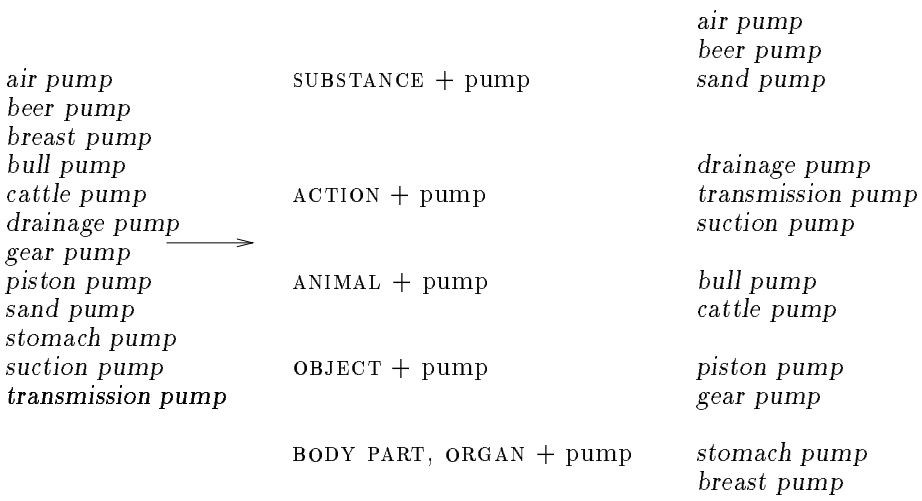

Figure 6: Structuration of NN terms according to the class of the modifier

This partition is a first step which helps to make the semantic facets of the head noun visible. Some associations are easily derived from what we know of the word *pump*: it is an agentive deverbal, and it belongs to the class INSTRUMENTALITY.

Consequently, we find in the list compounds exhibiting:

- the telic role of the noun:

  - SUBSTANCE + *pump* → predicate: *pump*(instrument: *pump*, theme: SUBSTANCE)

  e.g. *air pump* → (predicate: *pump*(instrument: *pump*, theme: *air*))

  - ACTION + *pump* → (predicate: ACTION(instrument: *pump*))

  e.g. *drainage pump* → predicate: *drain*(instrument: *pump*)

- the constitutive role of the noun

  - OBJECT + *pump* → predicate: *made-of*(object: *pump*, part: OBJECT)

  e.g. *gear pump* → predicate: *made-of*(object: *pump*, part: *gear*)

These patterns are predicted and interpreted by our set of rules. Other types of associations, too specific to be taken into account by our model, appear in the list: ANIMAL + *pump* (*cattle pump*) and ORGAN + pump (*stomach pump*), in which the missing predicates are respectively *feed* - i.e. *pump food for* - and *clean* - i.e. *pump the contents of*. We see that the underlying relation is more complex, because it includes also an implicit argument (*food*, *contents*) of the predicate. These are typically the specific patterns we would get from corpus observations, provided that they occur frequently enough to be taken into account by statistical analyses.

Conclusion

In this paper, we describe the implementation of a model that applies linguistic rules for the interpretation of nominal compounds in English, independently of any domain-knowledge. Our model exhibits the different kinds of relations that may be found in compounds:

- explicit vs implicit relation
- functional vs non-functional relation

and how these relational possibilities are related to morpho-syntactic and semantic lexical information.

We have experimented this model on a corpus of compounds; what comes out from this test is that the definition of general interpretation rules allows us to handle the semantics of nominal compounds, provided that we tolerate a certain amount of under-determination and multi-generation.

The implementation of this model shows that it is particularly difficult to handle:

- abstract nouns,
- non-functional relations and in particular general relations such as *part-of*, *subclass*, *identity*.
- disambiguisation of multiple interpretations.

Consequently, further semantic information must be acquired to refine this model, on the basis of the principles we have exposed. These general rules help to guide the retrieval of semantic clues for the interpretation of compounds in texts, through the identification of explicit relations in the context or through statistical observations. Such a general model is indeed, according to us, a necessary stage in order to guide the retrieval of semantic information from corpora and to account for less regular semantic patterns in the texts. This is the hypothesis we are currently implementing. The combination of general principles and corpus-based inferences should enable us to semantically structure the compounds of a given corpus, in order to improve text indexing.

References

[AB95] Houssem Assadi and Didier Bourigault. Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation de connaissances. In *Proc. of: 3èmes journées internationales d'analyse statistique de données textuelles*, Rome, Italy, December 1995.

[AB96] Houssem Assadi and Didier Bourigault. Acquisition et modélisation de connaissances à partir de textes : outils informatiques

et éléments méthodologiques. In *Proc of: 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, Rennes, France, January 1996.

[Bau79] Laurie Bauer. On the Need for Pragmatics in the Study of Nominal Compounding. *Journal of Pragmatics*, 3:45–50, 1979.

[CB95] Ann Copestake and Ted Briscoe. Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, 12(1):15–67, 1995.

[Don82] D.B. Mc Donald. *Understanding Compounds Nouns.* PhD thesis, Carnegie Mellon University, 1982.

[Dow77] Pamela Downing. On the Creation and Use of English Compound Nouns. *Language*, 53(4):810–842, April 1977.

[Fin80] Timothy Wilking Finin. The Semantic Interpretation of Nominal Compounds. In *Proc. of: First conference of AI*, 1980.

[FS94] Cécile Fabre and Pascale Sébillot. Interprétation sémantique des composés nominaux anglais et français. In *Proc. of: Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, Genève, Suisse, 1994.

[FS96] Cécile Fabre and Pascale Sébillot. Interprétation automatique des composés nominaux anglais hors domaine : quelles solutions ? In *Proc of: 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, Rennes, France, January 1996.

[Gri91] Jane Grimshaw. *Argument Structure.* MIT Press, Cambridge, 1991.

[LD95] Mark Lauer and Mark Dras. A Probabilistic Model of Compound Nouns. In *Proc of: Seventh Joint Australian Conference on Artificial Intelligence*, 1995.

[Lie83] Rochelle Lieber. Argument Linking and Compounds in English. *Linguistic Inquiry*, 14(2):251–285, 1983.

[Mar84] Elaine Marsh. A Computational Analysis of Complex Noun Phrases in Navy Messages. In *Proc. of COLING-84*, pages 505–508, 1984.

[MBF+90] George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five Papers on WordNet. Technical Report CSL 43, Cognitive Science Laboratory, Princeton University, July 1990.

[Mel84] Igor Melcuk. Un nouveau type de dictionnaire : le dictionnaire explicatif et combinatoire du français contemporain. In Melcuk et al., editor, *Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques I.* Presses de l'Université de Montréal, 1984.

[PAB93] James Pustejovsky, Peter Anick, and Sabine Bergler. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics*, 19(2):331–358, 1993.

[Pus91] James Pustejovsky. The Generative Lexicon. *Computational Linguistics*, 17(4), 1991.

[Res93] Philip S. Resnik. *Selection and Information : a Class-Based Approach to Lexical Relationships.* PhD thesis, University of Pennsylvania, 1993.

[Ryd94] Mary Ellen Ryder. *Ordered Chaos: the Interpretation of English Noun-Noun Compounds.* University of California Press, 1994.

[Seb93] Pascale Sebillot. Sémantique des composés anglais : approche générative, limites et applications. In *Proc. of: "Informatique et Langue Naturelle", ILN'93*, 1993.

[Sel82] Elisabeth Selkirk. The Syntax of Words. *MIT Press*, 1982.

[Spr94] Richard Sproat. English Noun-Phrase Accent Prediction for Text-to-Speech. *Computer Speech and Language*, 8:79–94, 1994.

[SV94] Wilco G. Ter Stal and Paul E. Van Der Vet. Two-level Semantic Analysis of Compounds : A case study in linguistic engineering. In University of Groningen, editor, *Papers from the fourth CLIN meeting*, pages 163–177, The Netherlands, 1994.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399